Towards a Bayesian Approach to Robust Finding Correspondences in Multiple View Geometry Environments

Cristian Canton-Ferrer, Josep R. Casas, and Montse Pardàs [*]

Image Processing Group, Technical University of Catalonia,
Barcelona, Spain
{ccanton, josep, montse}@gps.tsc.upc.es

Abstract. This paper presents a new Bayesian approach to the problem of finding correspondences of moving objects in a multiple calibrated camera environment. Moving objects are detected and segmented in multiple cameras using a background learning technique. A Point Based Feature (PBF) of each foreground region is extracted, in our case, the top. This features will be the support to establish the correspondences. A reliable, efficient and fast computable distance, the *symmetric epipolar distance*, is proposed to measure the closeness of sets of points belonging to different views. Finally, matching the features from different cameras originating from the same object is achieved by selecting the most likely PBF in each view under a Bayesian framework. Results are provided showing the effectiveness of the proposed algorithm even in case of severe occlusions or with incorrectly segmented foreground regions.

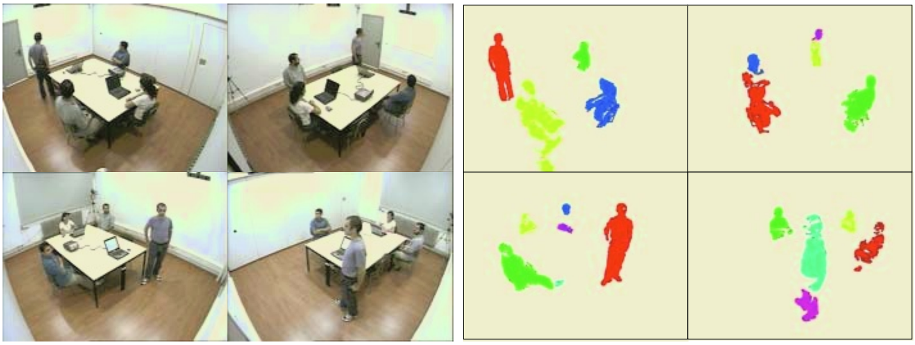
1 Introduction

Multi camera systems are being widely used for image and video analysis tasks in SmartRooms, surveillance, body analysis or computer graphics. Multiple view geometry has been addressed in [4, 9] from a mathematical viewpoint, but there is still work to be done for the efficient fussion of redundant camera views and its combination with image analysis techniques for object detection and tracking. In this framework, the current paper addresses the problem of finding meaningful correspondences between regions in different views.

In multiple camera environments, given an image sequence, the problem is to find the correspondences among feature points across the images that represent the projection of the same object in the real world in different camera views. Once these correspondences among all feature points are available, they can be used for object tracking and identification, motion analysis and scene analysis.

[*] This material is based upon work partially supported by the IST programme of the EU through the NoE IST-2000-32795 SCHEMA and IP IST-2004-506909 CHIL and by TEC2004-01914 project of the Spanish Government.

V.S. Sunderam et al. (Eds.): ICCS 2005, LNCS 3515, pp. 281–289, 2005.



(a) (b)

Fig. 1. In (a) there is an example frame from a multiview camera environment: SmartRoom at UPC. In (b) the result of a foreground segmentation and extraction of labeled blobs

The output of a multiview correspondence algorithm is a set of links, ideally representing how a unique point or an object in the real world corresponds to a specific position in each projected image. For applications like particle tracking or 3D tracking of a number of objects in a cluttered scene with strong occlusions, the redundancy in the different projections should help to overcome occlusion problems.

For a given frame in the video sequence, a set of N images are obtained from the N cameras. Each camera is modeled using a pinhole camera model based on perspective projection. Accurate camera calibration information is available. We use a segmentation algorithm by [10] based on Stauffer-Grimson's background learning and substraction technique [12]. In spite of the good performance of this technique, objects in the scene are oversegmented due to ilumination changes, reflexions and other distorting elements typical in non-controlled environments. Nevertheless, our algorithm proposes solutions to overcome oversegmentation problems and still finds accurate correspondencies.

Once foreground regions are extracted, a labeling process is performed based on connectivity rules, thus obtaining a set of enumerated disjoint foreground regions. Let us denote $\mathcal{B}^i_k(x,y)$, the set of pixels belonging to the k-th labeled foreground region at the i-th camera view and denote this set as blob. The work presented in this paper addresses the problem of establishing the correspondences among the set of blobs representing the projection of the same object in the different cameras. Once these correspondences are decided, this information can be used by further video processing modules such as the initialization of a multiocular tracker [6, 11], body gesture analysis [8] or 3D reconstruction [3].

There are two contributions in this paper. First, the application of a fast and simple cross-view point-to-point distance based on epipolar geometry, the *symmetric epipolar distance*, for guessing possible correspondences among views.

Then, a probabilistic approach to select the right geometric correspondences among blobs present in different views. This method takes into account occlusions and solves them by exploiting redundant information in multiple views.

2 Multiple Camera Object Correspondence

In order to compute candidate 3D locations of objects, an algorithm that guesses which foreground regions from different cameras belong to the same object is required. This grouping of foreground regions is depicted as a correspondence guess. In order to tell valid correspondence guesses from invalid guesses, a Point Based Feature (PBF) from each foreground region is extracted. A cross-view point-to-point distance is introduced to evaluate the validity of each correspondence guess. Correspondence hypotesis are generated by a sequential minimization procedure and, finally, a Bayesian decision algorithm evaluates the correspondence candidates in order to decide whether they are valid or not.

Existing correspondence algorithms [6, 11] rely on nonlinear least-square solutions or overdetermined linear systems. However, by increasing the views (cameras) or the number of detected objects, these systems turn out to be increasingly computationally expensive and, therefore, not appropriate for real-time systems. Other methods [1, 9] rely on homographies between images but, for wide angle lense cameras (such as the SmartRoom's case) homographies present significant distortion.

2.1 Symmetric Epipolar Distance

In order to find correspondences among blobs in different views, a distance that measures how close are the points of a set of points $\mathcal{D}(p) = \{\mathbf{x}^0, \mathbf{x}^1, ..., \mathbf{x}^{p-1}\}$ in p different views must be defined. In this case, closeness is understood as how well these points fit as projections of the same 3D point. This distance should be accurate, reliable and fast to compute. Under these conditions, *symmetric epipolar distance*, a ℓ_2-distance between points in different views based on epipolar geometry is employed. This distance was used in [14] to compare fundamental matrix estimations and has proven to be effective for our purposes. This distance is presented for the case of two points in two different views and afterwards extended to the general case.

Let $l\left(\mathbf{x}^i, j\right)$ be the epipolar line generated by the point $\mathbf{x}^i$ onto view j. We define the symmetric epipolar distance between two points $d_{\mathcal{SE}}(\mathbf{x}^i, \mathbf{x}^j)$, in the two views i,j, as:

$$d_{\mathcal{SE}}(\mathbf{x}^i, \mathbf{x}^j) \triangleq \sqrt{d^2(l(\mathbf{x}^i, j), \mathbf{x}^j) + d^2(l(\mathbf{x}^j, i), \mathbf{x}^i)}, \tag{1}$$

where $d(l(\mathbf{x}^i, j), \mathbf{x}^j)$ is defined as the Euclidean distance between the epipolar line $l\left(\mathbf{x}^i, j\right)$ and the point $\mathbf{x}^j$ as depicted in Fig.2. It can be shown that the extension of the symmetric epipolar distance for $p \geq 2$ points (in p different views) $d_{\mathcal{SE}}(\mathbf{x}^0, \mathbf{x}^1, ..., \mathbf{x}^{p-1})$ can be written in terms of the distance defined in Eq.1 as:

$$d_{\mathcal{SE}}(\mathbf{x}^0, \mathbf{x}^1, ..., \mathbf{x}^{p-1}) \equiv d_{\mathcal{SE}}(\mathcal{D}(p)) = \sum_{i=0}^{p-2} \sum_{j=i+1}^{p-1} d_{\mathcal{SE}}(\mathbf{x}^i, \mathbf{x}^j). \tag{2}$$

Hence, if the set $\mathcal{D}(p)$ is formed by projections of the same location in the 3D world, the distance between them should be very small (ideally zero).

The main advantage of this distance is its simplicity in contraposition with other distances where inverse matrix computations or homography estimation are required [9, 11]. However, despite this distance does not give a meaningful measure of a physical distance, it turns out to deal successfully with multiple view ($p \geq 3$) correspondence problems as explained in Sect. 2.3. When $p < 3$, two 3D points lying in the same epipolar plane would result in a small distance, hence the necessity of $p \geq 3$ in order to resolve the ambiguity.

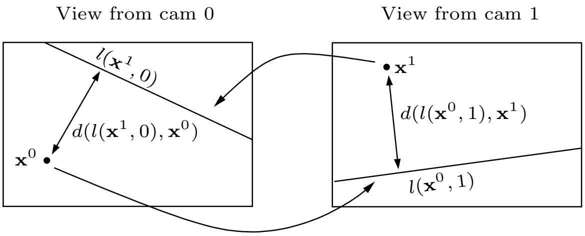


Fig. 2. Symmetric epipolar distance between two points $d(\mathbf{x}^0, \mathbf{x}^1)$

2.2 Point Based Features (PBF)

Point based features are those that only take into account a 3D point and the projection of this point in the existing views. To establish correspondences among blobs in different views, the centroids of blobs have been extensively used as a feature [6, 7, 11]. In the case of SmartRooms and cluttered scenes in general, blobs are affected by severe oclusions (due to furniture or other moving objects), hence their centroids are biased and not suitable as a correspondence feature. The tops of blobs in each view defined as

$$\mathbf{t}_k^i = \frac{1}{2} \left(\min_x \min_y \mathcal{B}_k^i(x, y) + \max_x \min_y \mathcal{B}_k^i(x, y) \right), \tag{3}$$

with i and k being the blob and camera index respectively, have been used as our PBF. It should be noted that the number of PBFs detected in one view may be different from another view.

In contraposition with other features, blob tops represent a meaningful point of the object and once correct correspondences are found, it can be used for forecoming analysis modules of the system (for example, estimating the height of a person). However, it must be said that for cameras positioned in a very low position, tops can be slightly biased due to the lack of visibility of the real top.

2.3 Finding Correspondence Candidates

Once a distance is defined, a method for finding and evaluating candidate correspondences between PBFs in different views is straightforward. Furthermore, if correct point correspondences can be found, blob correspondences across views are derived as well.

Let us suppose that a 3D point $\mathbf{X}$, has projections $\mathbf{x}^i$ on all the N cameras of our system. Despite this assumption is quite strong, let us define a method to find the correspondence guess and, afterwards improve it to deal with less restrictive assumptions. Then, we must choose the set $\mathcal{D}^\star(N)$ that accomplishes

$$\mathcal{D}^\star(N) = \min_{\mathcal{D}(N)} d_{\mathcal{SE}}(\{\mathcal{D}(N)\}), \tag{4}$$

where $\{\mathcal{D}(N)\}$ is the set of all possible correspondence candidates, that is all PBF combinations across all views.This process can be acchieved by either browsing all the possible combinations or reducing the search space with a priori geometric or probabilistic constraints [13].

In the case where multiple targets are present, M correspondence guesses should be determined. In this case, a sequential minimization procedure has been proposed according to [2]. This process is done by selecting the set $\mathcal{D}^\star(N)$ that minimizes the distance at a given iteration and then discarding the N points chosen for the next step, thus decreasing the search space dimension. This process is executed until M sets $\mathcal{D}_k^\star(N)$, $1 \leq k \leq M$, are selected. Results for this technique applied to find the correspondences between $N = 4$ views with $M = 4$ moving objects in a SmartRoom are shown in Fig.3.

In contraposition with the former method, a global minimization method to find the set $\mathcal{D}_k^\star(N)$ fulfilling the criteria

$$\mathcal{D}_k^\star(N) = \min_{\mathcal{D}_k^\star(N)} \sum_{k=0}^{M-1} d_{\mathcal{SE}}(\mathcal{D}_k^\star(N)), \tag{5}$$

has also been studied. That is to find the M sets $\mathcal{D}_k^\star(N)$ that minimize the total distance. Unfortunately, this problem turns out to be analogous to the overcomplete representation problem [5] and therefore an intractable NP-hard problem.

In the case when the 3D feature $\mathbf{X}$ does not have a projection in one or some of the views, our algorithm can still produce correspondence guesses if there are at list 3 projections of $\mathbf{X}$ available. We can find sets $\mathcal{D}^\star(p)$, $3 \leq p \leq N$ by allowing our minimization algorithm to build up the correspondence candidates $\mathcal{D}(p)$ including a null-projection in one or some views. Once a correspondence $\mathcal{D}^\star(p)$ is found, an estimation $\tilde{\mathbf{X}}$ of the 3D position of feature $\mathbf{X}$ can be estimated with a joint estimation back-projected ray with outlier rejection method [4]. By projecting back $\tilde{\mathbf{X}}$ onto all the camera planes we can obtain the set $\bar{\mathcal{D}}^\star(N)$ that completes the missing projections of $\mathcal{D}^\star(p)$ in the case where $p < N$. Moreover, since the information contributed by all the views is taken into account in the estimation $\tilde{\mathbf{X}}$, this new set $\bar{\mathcal{D}}^\star(N)$ refines the original one diminishing local errors.

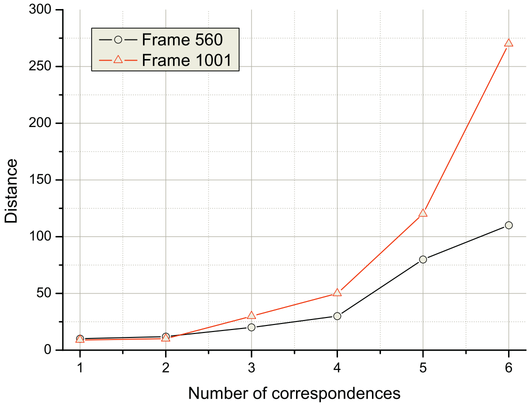


Fig. 3. Distance vs. Number of points selected with $N = 4$ and $M = 4$. The first 4 correspondences present a small and similar distance but, since the fifth corresponcence and following do not belong to any 3D point, their distances increase noticeably

2.4 Bayesian Correspondence Algorithm

The method proposed in the former subsection is able to produce M correspondence guesses for a given number of cameras p. However, since these parameters have to be selected manually, an adaptive algorithm that guesses the number of existing valid correspondences M and selects the optimum number of views p for each correspondence is required.

As a first step of this algorithm, it must be determined whether a correspondence $\mathcal{D}^{\star}(p)$ is formed by projections of the same 3D feature, denoting this property as belonging to the set $\mathcal{V}$, namely $\mathcal{D}^{\star}(p) \in \mathcal{V}$. For this task a Bayesian approach is employed. The posteriori probability of $\mathcal{D}^{\star}(p) \in \mathcal{V}$ given its symmetric epipolar distance $d_{\mathcal{SE}}(\mathcal{D}^{\star}(p))$ is formally:

$$P(\mathcal{D}^{\star}(p) \in \mathcal{V} | d_{\mathcal{SE}}(\mathcal{D}^{\star}(p)) = \frac{P(d_{\mathcal{SE}}(\mathcal{D}^{\star}(p)|\mathcal{D}^{\star}(p) \in \mathcal{V})P(\mathcal{D}^{\star}(p) \in \mathcal{V})}{P(d_{\mathcal{SE}}(\mathcal{D}^{\star}(p))}. \quad (6)$$

Since the priors $P(\mathcal{D}^{\star}(p) \in \mathcal{V})$ and $P(d_{\mathcal{SE}}(\mathcal{D}^{\star}(p))$ are wide and uninformative, Eq.6 can be rewritten as:

$$P(\mathcal{D}^{\star}(p) \in \mathcal{V} | d_{\mathcal{SE}}(\mathcal{D}^{\star}(p)) \propto P(d_{\mathcal{SE}}(\mathcal{D}^{\star}(p)|\mathcal{D}^{\star}(p) \in \mathcal{V}), \quad (7)$$

where the probability $P(d_{\mathcal{SE}}(\mathcal{D}^{\star}(p)|\mathcal{D}^{\star}(p) \in \mathcal{V})$ is modeled as a Gaussian distribution $\mathcal{N}(m_p, \sigma_p)$. Parameters m_p and σ_p are tunable depending on the accuracy of the segmentation algorithms and the geometry of the environment. This assumption is indeed valid after our observation of the average distribution of the distance $d_{\mathcal{SE}}(\mathcal{D}^{\star}(p))$. For each value of p, this distribution will be different, obbeying the rule $m_p \geq m_{p-1}$ since the less views are taken into account for the correspondence of the projection of $\mathbf{X}$, the less the overall distance. Conversely, the less the views employed, the less accurate the estimation $\tilde{\mathbf{X}}$ results. Hence, correspondences with the most views will be preferred. As a direct aplication of this theory, the implementation of our algorithm is based on the sequential selection of the sets $\mathcal{D}^\star(p)$ with larger p and higher probability in decreasing order. The algorithm stops when the probability is under a certain threshold α (for our experiments $\alpha > 0.7$).

3 Results

In order to evaluate the performance of the proposed algorithm, we used it to set correspondences in a SmartRoom equiped with 4 fully calibrated cameras where foreground blobs represented people. A sequence of 2000 frames was recorded simulating a presentation meeting with 4 atendees.

The images were segmented but, due to oversegmentation, we obtained more blobs per image than the number of persons. Consequently, the possible combinations grew up to an average of 20000 combinations per set of images. Even in this harsh conditions, the algorithm was able to perform effectively. For comparison purposes, Table 1 shows the performance of our algorithm for two types of PBFs, tops and centroids, where tops outperformed the existing algorithms based on centroids. Fig. 4 illustrates successful correspondences established among blobs representing people in a SmartRoom.

Table 1. Results of correct match and confidence ($\bar{P}$) of the correspondence estimation obtained by applying our correspondence algorithm on 200 frames chosen at random over a sequence of 2000

	Correct Match (%) PBF top	$\bar{P}$ (%) PBF top	Correct Match(%) PBF centroid	$\bar{P}$ (%) PBF centroid
Person 1	99.8	99.2	11.76	21.3
Person 2	99.77	98.62	41.18	43.27
Person 3	96.04	91.3	50.78	37.5
Person 4	95.8	85.6	44.31	29.0

4 Conclusions and Future Work

We have presented an algorithm for finding correspondences between regions in a multiple camera environment. The algorithm employs a probabilistic criterion for matching PBFs in different cameras, in our case the top of the blobs, and has proven very reliable, outperforming the existing approaches to this problem. The results of the algorithm are interesting for the initialization a multiocular tracking of multiple people system, multiview face or body detection, analysis and recognition.

Experimental results verify that we can obtain good quantitative 3D parameters from 2D image observations of people in the scene. We have demonstrated

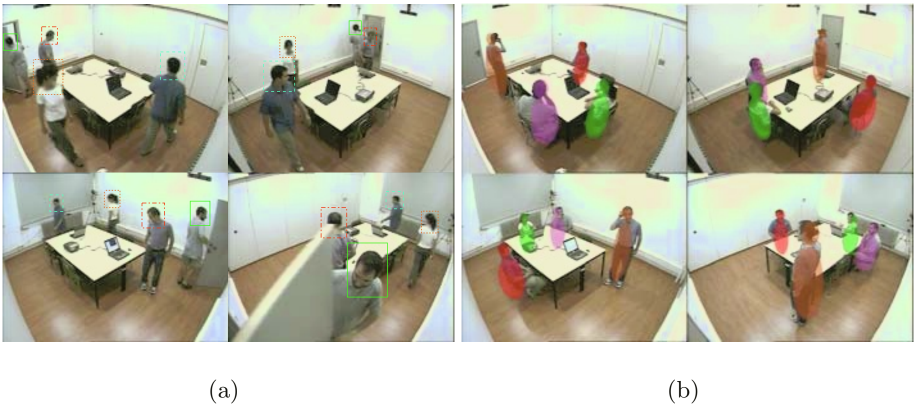


Fig. 4. Two examples of the application of our correspondence algorithm. Figure (a) shows its utility for multi-view face detection where the faces of the four people are being localized in all views providing reliable data for higher level analysis. In (b) correspondences are being correctly established in the framework of body analysis

that correspondences between blobs belonging to the same physical object can be found even if the object does not have a projection in some of the views. Future research perspectives involve the development of robust 2D/3D tracking of this PBF correspondences.

References

1. Black, J., Ellis, T.: Multi Camera Image Tracking. In *Proc. IEEE Work. on Performance Evaluation of Tracking and Surveillance*, 2001.
2. Boyd, S., Vandenberghe, L.: Convex Optimization. 1st edn. Cambridge University Press, 2004.
3. Eisert, P., Steinbach, E., Girod, B.: Multi-hypothesis, volumetric reconstruction of 3D objects from multiple calibrated camera views. In *Proc. IEEE Int. Conf. on Acoustics Speech and Signal Processing*, pp. 3509–3512, 1999.
4. Faugeras, O., Luong, Q.T.: The geometry of multiple views. 1st edn. MIT Press, 2001.
5. Figueras, R.M.: Image Coding with Matching Pursuit. MS Thesis, EPFL, 2000.
6. Focken, D., Stiefelhagen, R.: Towards vision-based 3D people tracking in a smart room. In *Proc. IEEE Int. Conf. on Multimodal Interfaces*, pp. 400–405, 2002.
7. Fuentes, L.M., Velastin, S.A.: People Tracking in Surveillance Applications. In *Proc. IEEE Work. on Performance Evaluation of Tracking and Surveillance*, 2001.
8. Gavrila, D. M., Davis, L. S.: 3D Model Based tracking of humans in action: a multi-view approach. In *Proc. of Computer Vision and Pattern Recognition*, pp. 73–80, 1996.
9. Hartley, R.I., Zisserman, A.: Multiple View Geometry in Computer Vision. 2nd edn. Cambridge University Press, 2004.

10. Landabaso, J.L., Xu, L.Q., Pardàs, M.: Robust Tracking and Object Classification Towards Automated Video Surveillance. In *Proc. Int. Conf. on Image Analysis and Recognition*, pp. 463–470, 2004.
11. Mikic, I., Santini, S., Jain, R.: Tracking Objects in 3D using Multiple Camera Views. In *Proc. Asian Conf. on Computer Vision*, pp. 234–239, 2000.
12. Stauffer, C., Grimson, W.E.L.: Adaptive background mixture models for real-time tracking. In *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition*, pp. 252–259, 1999.
13. Triggs, B.: Joint Feature Distributions for Image Correspondence. In *Proc. IEEE Int. Conf. on Computer Vision*, pp. 201–208, 2001.
14. Zhang, Z.: Determining the epipolar geometry and its uncertainty-a review. In *Int. Jour. of Computer Vision*, 27(2):161–195,1998.